DATA ANALYSIS

Welcome back.
So today we're going to be talking about part three, data visualization.
Remember, there's kind of three pieces of becoming a data scientist that you really should be working towards mastery in.
Data organization, visualization, and data analysis.
So just like with data organization and what's coming up in data analysis, we're dividing data visualization up in kind of three levels.
So, the first level is kind of the typical stuff that you're going to be seeing again, using Excel, Power BI.
We're generating these sorts of really kind of classic goal data presentations.
So, you're talking about trending, and histograms and trend lines and kind of this level that's some pretty basic statistics plus nice visual graphics that most everybody can understand and has seen before.
Ideally, with all of these final types of presentations, the thing about them is you can make the most amazing graphic, but if it's something that somebody can't understand,
that's really not very useful, and so just keep that in mind.
So, there's some really beautiful things that you can do with data visualization, and sometimes it does not make the science clear.
So even though you really want kind of this artistic background with your data presentation.
Make sure that it's clearly illustrating what you need it to illustrate for your audience, for the presentation side.
So, let's jump into level one, data science or data visualization.
And then kind of talk about what that means.
And then we will move into kind of the level two and the level three stuff.
And data visualization and data analysis kind of have this interesting jump that we will talk about later in your level three stuff.
And just to kind of give you a primer to what we'll talk about,
there's this trend in data science to
move away from kind of this classical statistics and
kind of model-based approach to this machine learning approach.
And both level two or level three stuff in visualization, as well as
in analysis, kind of we'll talk about that a little more and
kind of why that trend is occurring and what it means.
So cool.
So again, visualization, kind of there's two pieces of visualization.
First is you understanding the data,
and then the second piece is explaining results to other people.
And so that is kind of why in the hierarchy of events,
we have organize the data and then visualize.
Visualize the data.
And then analyze the data.
Because you really wanna understand what you're trying to do

with your data.
Prior to running an analysis on your data.
So getting that data intuition is really key and
a really easy way to get initial data.
Kind of.
Understanding is to visualize it.
So.
So here's an example of a visualization.
So maybe you have a bunch of different data points, and they're
just thousands of data points, and you don't know what they are.
So, there's some kin of trends in data science that you want to be
able to predict behaviors based on some sort of dynamic of the data.
And so a very common one that you hear about is a normal distribution.
And these are these nice bell shaped curves and a lot of you have seen
gas in distribution that sort of thing, but with just looking at
a data set you have no idea if it's gonna be filling a gas in curve.
Or may look like something like this,
where you have kind of these two separate homes.
And so you wanna be able to see that in some manner.
So whether or not you can, or one of those people can just visualize it
yourself, which is a very, very, very small percentage of us.
But most of us need some sort of representation.
So the histogram is a super common one.
Basically is says, how many times did a certain event occur?
So, how many students got A's, B's, C's, or D's?
Or, kind of maybe
some different bins you might want to put something in.
Actually, in education that's a very common thing is to see these
bi-mobile distributions where you have kind of one peak.
At one location kind of an A, B range and
then another P down at the D, F range.
At least in like, the high school level.
And there's a bunch of different reasons for that but, a lot of it's
just student engagement and things that are going on in students lives.
So, you get these interesting distributions.
So you wanna be able to see that.
And so, before you analyze your data.
Because you can't do all the same things,
if the data are not normally distributed.
You can't run T tests very easily.
You can't do some of the things where you're saying, hey,
you know this is three standard deviations away from the mean.
So it's something that's probable to happen or not to happen.
If you're three standard deviations away from the mean, it's 99% chance
that it's not related to that mean or is a pretty far out layer, so.
So understand that is super important.

And kind of knowing what your data will look like in knowing what sort of tools you can use are, absolutely critical. So.
The second piece is
the more classical presentation sort of thing.
So an example, you know,
there's some great tools out there right now.
So Excel has, continues to make amazing changes,
where you can basically say, hey, let's make a chart of x, y, or z.
And it'll just do it for you and be like okay, and
this is a beautiful chart.
I just pulled up Excel's chart maker and this is just
a screenshot of the tutorial that they have on the beginning of it.
So you can just do this stuff and it'll make these amazing charts.
And you know, we look at them and we can kind of understand what they're doing, but one think I really wanna caution people against as you become a data scientist, is not to just blindly trust these tools.
And you really wanna understand what these tools are doing behind the scenes so, that you can kind of understand and explain to somebody when you say hey, I'm presenting this data, you know maybe you're trying to make some sort of decision for your company.
You don't want to just be like, hey this tool just spout out this really nice graphic for me, and here it is.
You want to be able to explain what's going on behind there.
And that'll allow you to kind of really manipulate that data.
So, as said, there's some great tools for
this kind of level one stuff.
So the ones that we talk about through this course are PowerBI and Excel.
Tabloid has some great things.
Even some of the programming languages, Spark,
sorry, Python and R both have some great ways to visualize data.
So use those, those are some of the tools that you're gonna be wanting to understand and be able to feel pretty comfortable with.
But I really want to re-emphasize this morning, before you go and present any type of data, you really wanna understand those tools that you're doing and use to get those data into a presentable form.
Because there's nothing worse than going into a presentation and grilling the presenter about their data because,
you see some sort of trend in there.
And them being like,
I don't really understand how I got this information, right?
Like,you don't wanna be that person as the presenter,
understand what's going on.
So, but as a level one person.
You know, taking this idea and
looking at your data and starting to get that data intuition.

So like doing the previsualization and then understanding what you're doing as kind of your final presentation is absolutely critical.
So level two, In level two, you're gonna be taking data types that are a little bit different so, here's an example of some no cost string so basically this idea would be something like Facebook, right.
So in Facebook, you have a bunch of friends and maybe those friends have a certain amount of messages that they send back of forth.
They're gonna be closely clustered to you because,
you're You're reading lots of messages from them.
But there's a few friends out there that you have
that maybe are further away.
You maybe only send a message to once a month, or once a year.
They're still connected to you, but they have a lower weight,
in this sense.
And so, you can start generating these graphics that kind of represent that.
And again, you wanna be able to explain, what this complicated relationships are in a way that, with a short explanation or
sort of interaction with your end user or
even a written description for that graphic.
They can understand, too, what you're trying to get across.
And so the level two stuff is gonna be taking data that are maybe not that kind of classical, automatically generated sort of
formats from some of these These other tools.
But maybe some more complicated structures where you'll be, okay,
I want to visualize this.
Maybe this three-dimensional or
four-dimensional space, such that somebody else can understand it.
And so, it's kind of more on the analytical side of things.
So, you can really make some amazing graphics from,
if you're in that level two space, and
these graphics are going to be these design, decision guiding graphics.
So, graphics that are going to take the user,
suck them in and really give them a good grip on some
data that they might not have had a understanding that before hand.
This includes things like interactive designs and of course,
you're gonna still want to know how to apply core analytical methods to this data.
So that's kind of a level two structure.
Again, PowerBI and Excel do some of these things for you, but
at this point you're gonna be moving into doing stuff with Python, or D3.js, or R.
And, I think in the future, it's going to be really interesting,
because these data sets are getting more complicated, and
you're getting these multidimensional data sets.
And what I mean by multidimensional,

is these are data that have a whole bunch of different types of columns.
So, every column in a data set, you can think of as a dimension that
you're gonna look at, that you could investigate, right.
You could compare, seat number on a car to car weight,
right, those are two dimensions that you could graph on a x and y, and
so you can describe as an x and y, you can consider a dimension.
And so how do you visualise that when you
have 500 things about a car, in a way that kind of
allows somebody to explore that information.
So virtual reality, I think, is gonna be an interesting tool and
people are gonna start to use it.
Is where you can kind of walk into your data and
start to manipulate it.
But we'll see.
That's just my speculation.
But I think that's still fits into this kind of level two structure.
So.
And as I mentioned before, level three is going to be sort of a jump.
And so, right now we're kind of dealing with kind these classical
columnar sorts of data.
Were you're dealing with data sets that, you know,
kind of make sense of Level three visualization is going to be this
ability to present architectures of experimental designs and
how you're gonna analyze the data.
Because, as I said, data science is also bringing this huge
component from machine learning, and machine learning you know,
it's not always about, you know,
kind of being able to visualize that data from the top down.
Because you're gonna let, let the,
the machine do a bunch of these experiments and hypotheses.
And so you want to be able to tell your end user basically, how are you
setting up those, those parameters for the machine to learn with.
And so my view of, of level three data is basically being
able to explain clearly what sort of manipulations you're gonna let
the computer do, such that it can analyze the data for you.
Because then you can have a discussion about what those
results are.
So if you don't have a clear understanding of how you set up your
system, and they don't have a clear understanding of how you
set up your system, those results really aren't that meaningful,
because you don't know how to communicate about them.
Because you're like well, yeah,
it predicted what we expected it to predict and it's been really good.
It has like an 85% or 95% accuracy rate but,
I don't know how it got there right?
So again you don't wanna be that person, you don't wanna be like hey

this is working really well but, I have no idea how it works, so.
Thanks.
So, neural nets are a structure for
allowing computers to make decisions.
So, in these input neurons, pretend that we're visualizing a spider and
so input neuron, that top one on the top left might be the color black.
Is it black?
And so it's gonna be okay, so it's either black or not.
So that's gonna be either on or off.
And then the next thing is, is it shiny?
And that's gonna be on or off.
The next thing might be, does it have eight legs?
Thanks.
Thanks.
Thanks.
And then, depending on what Inputs are actually firing, those are going
to hit these integration neurons and kinda make decisions based on that.
Thanks.
And so that might feed into the integration neurons two and four.
And those, then, send out information to these output neurons.
Maybe you leave it alone.
Or you run away screaming.
So.
Depending on what it is. Right.
And so
you kind of want to be able to describe this.
How this decision making process is going to occur and then you're
gonna take a data set, or data sample potentially, and apply it.
And say hey,
every time somebody saw a black widow, they ran away screaming.
Let's let the machine figure out how to figure that out.
So, So, that's kind of where I kind of see level three visualization.
Like, the data that comes back after you've machine learned something,
can certainly be in a format that you'd apply level two and
level one techniques to, where you're basically saying, hey,
these are the results.
But to really have a conversation about those results,
I think you really need to be able to clearly visualize and
make visual graphics that understand and explain to your user and
to yourself, how your system is set up?
Thanks.
So kind of the tools that you have for level three data, you wanna
be able to present things all the ways that you can in level two.
But you also really wanna be able to present machine architectures and
machine learning architectures.
So that's kind of where I see level three.

So that completes our data visualization chapter.
And next we're gonna talk about data analysis,
and in that chapter we'll talk about just how to deal with all these different pieces of data that we're getting in.
We'll, move on to kind of my view of how to become a data scientist and some of the paths that you can take to becoming a data scientist.
And kind of reiterate all the pieces that we've talked about today.
Thanks.